The Social World of Twitter: Topics, Geography, and Emotions

Daniele Quercia[§], **Licia Capra**[‡], **Jon Crowcroft**[§]
[§]The Computer Laboratory, University of Cambridge, UK
[‡]Department of Computer Science, University College London, UK
dq209@cl.cam.ac.uk, l.capra@cs.ucl.ac.uk, jac22@cl.cam.ac.uk

Abstract

Debate is open as to whether social media communities resemble real-life communities, and to what extent. We contribute to this discussion by testing whether established sociological theories of real-life networks hold in Twitter. In particular, for 228,359 Twitter profiles, we compute *network metrics* (e.g., reciprocity, structural holes, simmelian ties) that the sociological literature has found to be related to parts of one's social world (i.e., to *topics*, *geography* and *emotions*), and test whether these real-life associations still hold in Twitter. We find that, much like individuals in real-life communities, social brokers (those who span structural holes) are opinion leaders who tweet about diverse topics, have geographically wide networks, and express not only positive but also negative emotions. Furthermore, Twitter users who express positive (negative) emotions cluster together, to the extent of having a correlation coefficient between one's emotions and those of friends as high as 0.45. Understanding Twitter's social dynamics does not only have theoretical implications for studies of social networks but also has practical implications, including the design of self-reflecting user interfaces that make people aware of their emotions, spam detection tools, and effective marketing campaigns.

1 Introduction

In 1983, political scientist Benedict Anderson published a book titled "Imagined Communities" in which he argued that the sense of community commonly referred to as citizenship "is imagined because the members of even the smallest nation will never know most of their fellow-members, meet them, or even hear of them, *yet* in the minds of each lives the image of their communion" (Anderson 1983). As we shall see in Section 2, the literature tends to find communities created by electronic means of communication (including Facebook and Twitter) similar to Anderson's imagined nation-state community (Anderson 1983) - members of electronic communities may have never met in person, yet to some degree regard themselves as part of a larger whole (Gruzd, Wellman, and Takhteyev 2011).

To quantitatively assess the extent to which social media communities resemble real-life ones, we test whether established sociological theories of real-life (offline) social networks still hold in Twitter. In so doing, we make four main contributions:

- We compile a list of the *network metrics* that the literature has found to be related to parts of one's social world (more specifically, related to *topics*, *geography* and *emotions*). These metrics include reciprocity, simmelian ties, and network constraint. We compute these metrics for 228,359 Twitter profiles we have crawled (Section 3).
- (On Topics). We classify the topics of 31.5M tweets and study the relationship between topical diversity and the previously computed network metrics (Section 4). We find that social brokers in Twitter are opinion leaders who take the risk of tweeting about different topics (influential Twitterers tend to be specialized in specific topics instead (Cha et al. 2010)).
- (On Geography). We geo-reference Twitter user-specified locations, compute a measure of geographic span (i.e., geographical dispersion of one's followers), and test its relation to network metrics (Section 5). We find that the majority of users have geographically local networks. Also, for each egonetwork, we consider four alternative versions (each with ties of increasing social strength) and learn that, the stronger the considered ties, the more geographically local the corresponding networks.
- (On Emotions). We determine the extent to which our tweets express emotions and test the relationship between network metrics and emotions (Section 6). We find that social brokers express not only positive but also negative emotions, and that users who express positive (negative) emotions strongly associate with each other.

We conclude by discussing how one could build practical applications upon this work (Section 7).

2 Related Work

The vast majority of empirical work on information advantage in networks is "content agnostic" (Hansen 1999) - the actual information flowing between connected individuals is rarely observed. Research has focused on network structure instead, and has consistently linked it to information advantage. However, as Burt writes, these studies bear limitations:

"The hubs in a social network were argued to have advantaged access to information and control over its distribution [...] However, the substance of advantage, information, is almost never observed [...] The next phase of work is to understand the information-arbitrage mechanisms by which people harvest the value buried in structural holes [...] More generally, the sociology of information will be central in the work [...]" (Burt 2005).

To date, few works go beyond structure. In their 2011 article "The Diversity-Bandwidth Tradeoff", Aral and Alstyne combined social network and performance data with direct observation of the information content flowing through e-mail communication and found that total volume of novel information increased not only with network size and network diversity (as one would expect), but also with frequency of communication. In other words, novel information is gathered both through a diverse network structure and from frequent communication (from what they called "thick edges").

To determine whether information content flowing between users is novel or not, researchers have extracted topics from it and considered topical diversity as a measure of novelty. Since emotions differ from topics but still reflect the exchanged information, Kivran-Swaine and Naaman studied how emotions are shared by 628 Twitter users (Kivran-Swaine and Naaman 2011). They found that those who express emotions tend to have more followers and sparser networks.

In addition to topics and expression of emotions in networks, researchers have considered the impact of geography on formation of ties. Only few months ago, Takhteyev, Gruzd, and Wellman studied the geographic distribution of 3K Twitter egonetworks and showed that, despite the seeming ease with which long distance ties can be formed, ties are constrained by distance and, as a result, most of them are geographically local (Takhteyev, Gruzd, and Wellman 2012).

From this brief literature review, one concludes that, despite some preliminary related work, we hitherto lack a detailed understanding of *how geographically-constrained Twitter users share information about diverse topics and express emotions under a variety of network conditions*. There has not been any study of how topics, geographical features, and expression of emotions are related to network structure for the same set of Twitter users. In this paper, we aim to close this gap: we do so by compiling a set of network metrics (Section 3) whose hypothesized associations with topical diversity (Section 4), geography (Section 5), and expression of emotions (Section 6) will then be verified.

3 Network Metrics

To begin with, one has to decide how to crawl the Twitter graph. Ideally, to obtain unbiased network metrics, one has to crawl either the complete graph (which the rate-limited Twitter API makes difficult) or individual egonetworks - in that case, for each ego, one would have the complete set of edges and, as such, the resulting network metrics will suffer from little bias. We opt for the second option and, to control for any variability in the use of language across geographic areas, we have preferentially chosen Twitter profiles from London, which has been chosen because the higher the adoption rate of a service, the lower the demographic bias, and London was the top Twitter-using city in the world until the beginning of 2010 (Butcher 2009). We chose three popular London-based seed profiles of news outlets: the free subway newspaper Metro, the center-left newspaper The Independent, and the tabloid The Sun. These news outlets cover the entire UK political spectrum and have high penetration rates in the city.

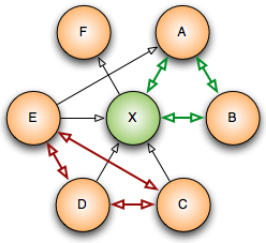


Figure 1: Topology of *X*'s egonetwork. Ties in green are *X*'s Simmelian ties.

Egonetworks. A user's *egonetwork* consists of the user ("ego"), the "alters" to whom the ego is linked, and the following/follower relationships between ego and alters and those among alters. In Figure 1, *X*'s egonetwork has six alters ($A - F$), *X* has a *unidirectional* relationship with *E* (X is followed by E) and a *bidirectional* relationship with *A* (X both follows A and is followed by A); furthermore, *E* has a undirectional relationship with *A*, which is totally independent of their relationships with *X*, yet $E \rightarrow A$ is still part of the egonetwork as it is between two of *X*'s alters.

Crawling. We crawled Twitter from 27 September to 30 December 2010 and selected users who have made their profiles publicly available (our crawling indicated that approximately 99.7% of users have done so) and have posted at least one hundred tweets (because of API limitation, we crawl up to 200 tweets for each user). In so doing, we have gathered: 258,895 profiles (for the 1,021 egos and 257,964 alters); 31,565,708 tweets for 240,982 users; and 10,254,969 network edges (follower relationships).

Validation of crawled data. One concern with data such as ours is that not all profiles necessarily belong to real people to which sociological norms could be expected to apply: some "users" might be businesses and organizations. To determine what profiles might not belong to real people, we crawled the TrstQuotient score provided by Infochimps.com. This score is in the range $[0, 100]$ and reflects the extent to which a user is "normal" (for example, based on its number of followers) - very low TrstQuotients are indicative of abusive or spam accounts. The resulting distribution of TrstQuotient values of the egos from our dataset is a skewed normal distribution ($\mu = 57.7, \sigma = 21.4$), which fortuitously indicates that the vast majority of egos in the dataset have a very high TrstQuotient value and are thus likely to be real users. We then filter

away profiles with very low TrstQuotient (those in the first quartile) and are left with 228,359 profiles and, still, 1,021 egos.

Four versions of each egonetworks. Twitter itself provides no information about the strength of a relationship other than the binary following/follower relationship. Whilst in many ways this is sufficient to determine an affinity between any given pair of users, in this study we also consider stronger ties by examining symmetric ties (reciprocal relationships) and exchange of `@replies` messages. Considering different types of ties might yield drastically different results in terms of resemblance of the Twitter network to real (offline) social networks. To test whether this is true, for each egonetwork, in addition to the original version, we build three other versions:

Reciprocal. This version considers only bidirectional edges - edges between pairs of users who follow each other. The resulting egonetworks are *undirected* graphs made of 24.1% of the original ones.

1-way Interaction. These egonetworks reflect stronger relationships between users by creating directional links from a user to another - *from* a user who sent an `@reply` *to* another user. These egonetworks include a total of 4.4% of the original set of edges.

2-way Interaction. Finally, *2-way interaction* networks are the most selective and consider only those users who have *exchanged* an `@reply` with each other. This filtering resulted in having just 0.6% of the original set of edges.

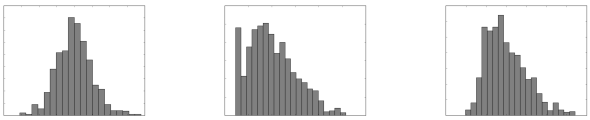
In filtering edges, our use of only the last 200 tweets for a user means that only 4.4% of the original edges were supported by an @reply and made it through to the filtering. This means that the conclusions drawn from the *1-way interaction* and *2-way interaction* networks cannot be considered as thorough as those drawn from the *original* and *reciprocal* networks. However, we will see that both 1-way and 2-way interaction networks offer insights that confirm what has been found in previous studies on the impact of tie strength in Twitter (Huberman, Romero, and Wu 2008).

Following, Followers, and Status. The simplest network metrics we consider are number of Twitter followers (in-degree), number of following (out-degree), and network status, which, in Twitter, is computed as the ratio between number of followers and number of following (Cha et al. 2010).

Reciprocity. The first egonetwork metric we consider is its reciprocity r, which is the proportion of its edges that are bidirectional (reciprocal). It ranges in $[0, 1]$ - high values correspond to socially closely-knit egonetworks, while low values correspond to linked users who each belong to different communities. Particularly low reciprocity values could be indicative of a celebrity's egonetwork (high in-degree, low out-degree network: many followers but few symmetric ties) or a spammer's egonetwork (low in-degree, high out-degree network: many friends but few symmetric ties). Reciprocity of our egonetworks is normally distributed (Figure 2a), with almost all egonetworks exhibiting roughly the same degree of reciprocity; on average, 22% of edges in an egonetwork are mutual ($\mu = 0.219$, $\sigma = 0.070$).

Simmelian Ties. Reciprocity is a measure that considers dyadic relationships. However, social scientists have consistently shown that also triadic relationships are important as they offer far greater insights into the connectedness of egonetworks (Tortoriello and Krackhardt 2010). To also consider triadic relationships, we examine those ties called "simmelian ties". These are, by definition, ties embedded in closed triples. In Figure 1, the triad in green ($X \leftrightarrow A \leftrightarrow B$) consists of Simmelian ties as it is a triad that includes the ego; by contrast, the triad in red ($C \leftrightarrow D \leftrightarrow E$) consists of edges that are not Simmelian (they are among alters and do not include the ego). In 1908, the sociologist Simmel argued that the fundamental building block of social relations is not the dyad but the triad. A dyadic relationship is quantitatively different from a relationship embedded in a group, and this difference cannot be explained only by the dyad's tie strength. To see why, consider the relationship between two partners who just met. This is a dyadic relationship and changes if, after a while, the couple will have a baby - then the relationship becomes a "simmelian tie", and the nature of the relation is best explained if it is considered to be embedded in the triadic relationship. David Krackhardt and colleagues have shown that simmelian ties matter for different reasons (Tortoriello and Krackhardt 2010): (a) cooperation - pairs of individuals are more likely to cooperate if their relationship is embedded in a triad; (b) tie decay - decay rate of simmelian ties is far slower than that of symmetric ties and that of asymmetric ties (the latter are the fastest to decay); (c) innovation - the more simmelian ties one has, the more productive (as per, for example, number of patents) one is. To paraphrase this research in the context of Twitter, for each egonetwork, we examine the proportion of its mutual ties that are simmelian. The corresponding frequency distribution is log-normal (Figure 2b) - many egos have no Simmelian ties, but those that do tend to have only a very small number of them.

Network Constraints. Reciprocity and proportion of Simmelian ties are both – in different ways – measuring how "mutual" (closely-tied) an egonetwork is. Next we consider the presence of *structural holes* in an egonetwork. In his book "Structural Holes: The Social Structure of Competition", sociologist Ron Burt put forward the idea that innovation is tied to structural empty spaces (structural holes) in networks. In knowledge-based workplaces, the highest-ranked ideas come from managers who have contacts outside their immediate work groups, and that is because their contacts span structural holes (gaps between discrete groups of people). By contrast, those whose contacts are all connected with one another have no access to structural holes and no opportunities to broker connections. Brokerage opportunities are computed with a measure called *network constraint*, which reflects the extent to which an ego's connections are concentrated on a clique of interconnected alters, meaning little access to structural holes (Burt 1992). Higher constraint means less brokerage opportunities, whilst lower constraint means more access to structural holes and brokerage opportunities. We compute



(a) Reciprocity (b) Simmelian Ties (c) Constraint

Figure 2: Three of the network metrics under study: **(a)** network reciprocity; **(b)** simmelian ties; and **(c)** network constraint. For reasons of space, these plots show the shape of the distributions; for the actual magnitude and values (e.g., mean, standard deviation), one should refer to the text.

network constraint as per Burt's original formulation (Burt 1992). The constraint between a pair of users (i, j) is:

$$c_{ij} = (p_{ij} + \sum_q p_{iq} p_{qj})^2 \times 100 \qquad q \neq i, j \qquad (1)$$

where p_{ij} is defined as the proportion of user i's "time and energy" spent on user j, and is measured as $p_{ij} = z_{ij} / \sum_q z_{iq}$.The notion of "time and energy" is thus specified as a weight z_{ij} between a pair of users, which could also be binary, like in our case where $z_{ij} = 1$ iff there exists an edge (i, j). After computing p_{ij} and c_{ij}, we compute the logarithm of the aggregate constraint C_i of i's network: $C_i = ln \sum_j c_{ij}$. The logarithm is used because the distribution of constraint is skewed. The resulting frequency distribution of the log-transformed constraint is a "skewed" normal distribution (Figure 2c).

Correlations of Network Measures. Having a variety of network metrics at hand, one might now wonder whether these metrics reflect very similar aspects or whether they reflect aspects that significantly differ from each other. One expects a strong positive correlation between reciprocity and proportion of Simmelian ties. That is because reciprocity is a proportion of reciprocal ties in the network and a triad (of Simmelian ties) consists of three reciprocal ties. Indeed, in our dataset, we find a significant correlation coefficient of $r = 0.58$ between reciprocity and proportion of Simmelian ties. By contrast, there is no correlation between any of the other pairs of network metrics - that is, there is no correlation between reciprocity and network constraint and between proportion of Simmelian ties and network constraint. We thus take all of these network metrics and to test their associations with topical diversity (Section 4), geographic span (Section 5), and expression of emotions (Section 6), so to shed more light into the research question we are attempting to answer.

4 Networks and Topical Diversity

Individuals with less constrained networks have greater access to structural holes and therefore greater brokerage opportunities (Burt 1992). The literature suggests that access to structural holes is highly beneficial in terms of exposure to diverse ideas: "people whose networks span structural holes have early access to diverse, often contradictory, information and interpretations, which gives them a competitive advantage in seeing good ideas [...] people connected across groups are more familiar with alternative ways of thinking and behaving." (Burt 2005). We posit that these "alternative" viewpoints manifest themselves on Twitter as egos tweeting on a diverse range of topics. A user who tweets mainly about politics would have a lower diversity than a user who is conversant in a more wide range of topics such as sport, entertainment and technology. Since we are interested in determining whether the diversity in a user's tweets can be used to determine their brokerage opportunities, the hypothesis we will test is:

Hypothesis 1 - Tweeters with higher diversity have higher brokerage opportunities.

Topical Analysis. To verify this hypothesis, the first step is to categorize each tweet into a set of discrete topics. We do so using three APIs. Had only a single API been used, bias might have been introduced into the pre-classification; the performance of one API may differ from that of another for certain topics, for example. For this reason, we have used these three APIs:

- *Alchemy API* (`http://www.alchemyapi.com/`) is a suite of natural language processing tools and is capable of assigning a plain English category to any given string of text (a tweet, for instance), along with a certainty score from 0.0 to 1.0, which represents the API's degree of belief that the text pertains to that category. Since Twitter users often tweet URLs relevant to the topics they discuss, we classify not only their tweets, but also the links they broadcast. *Alchemy* can choose from the following 12 topics: Arts_Entertainment, Business, Computer_Internet, Culture_Politics, Gaming, Health, Law_Crime, Recreation, Religion, Science_Technology, Sports, and Weather.
- *OpenCalais API* (`http://www.opencalais.com/`) comprises a suite of tools developed by Thompson Reuters that includes named entity extraction and text classification. When provided with a block of text, the API returns up to three topics, each with a belief score between 0.0 and 1.0. *OpenCalais*'s topics differ from *Alchemy*'s and are as follows: Business_Finance, Disaster_Accident, Education, Entertainment_Culture, Environment, Health_Medical_Pharma, Hospitality_Recreation, Human_Interest, Labor Law_Crime, Politics, Religion_Belief, Social_Issues, Sports, Technology_Internet, War_Conflict, and Weather.
- *Textwise SemanticHacker API* (`http://textwise.com/`) consists of tools for performing semantic analysis on bodies of text. It takes either a string of text or a URI (in which case, text is mined from the document defined by the URI), and returns a set of categories that pertain to it. As with *Alchemy* and *OpenCalais*, each these categories is paired with a belief score from 0.0 to 1.0. The main categories are eleven: Arts, Health, Science, Business, Home, Society, Computers, Recreation, Sports, Games, and Reference.

Topical Diversity. Having computed the topical distribution

	Network Constraint (no access to structural holes)				Reciprocity	Simmelian	Following	Followers	Status
	Original	Reciprocal	1-way Msg	2-way Msg					
Diversity	**-0.14**	**-0.22**	**-0.42**	**-0.48**	0.10	0.13	0.11	0.03	**0.21**

Table 1: Correlation coefficients r between network properties and topical diversity of tweets. Highlighted are those results that are statistically significant (p-values < 0.01).

of each individual tweet, we can now estimate an entire profile's topical diversity and do so by using the Shannon diversity theorem (entropy): $H' = -\sum_{i=1}^{S} p_i \ln p_i$, where S is the number of topics and p_i is the relative proportion of the i^{th} topic among the user's tweets. As our work has separately detailed (Quercia, Askham, and Crowcroft 2012), the values of topical diversity for the three APIs are strongly correlated at *profile* level (minimum correlation being $r = 0.94$) and, as such, we next report the results produced by the classification of *Alchemy* API only. The resulting distribution is power-law: most users have low diversity, while few users engage in discussions on a wide range of topics. That is unsurprising, as it is expected of individuals to primarily tweet about the subset of topics in which they are interested or knowledgeable.

Hypothesis testing. By computing the correlation coefficients between topical diversity and network measures (Table 1), we learn that users who tweet about diverse topics tend to:

- Have greater access to structural holes. The coefficients for all network constraints are negative, with a coefficient for the *2-way interaction* networks as low as -0.48. Figure 3a plots topical diversity against network constraint.
- Enjoy higher network status. Highly-diverse users are not characterized by a considerable number of followers or following, but by having higher network status (the number of followers is higher than number of following).

The hypothesized relationship between topical diversity and access to broker opportunities is confirmed, and that corroborates the theory that network diversity provides information advantage in part by providing access to diverse pools of expertise. As such, in Twitter, a basic premise of brokerage theory is supported: that disconnected network neighborhoods house dissimilar expertise and knowledge, which brokers tap into by reaching across structural holes.

5 Networks and Geographic Span

Previous studies such as "Imagining Twitter" (Gruzd, Wellman, and Takhteyev 2011) have questioned whether Twitter can be considered a community in the traditional sense, given that Twitter "friends" may never have even met in person and social media are increasingly reorganizing the society into so-called "imagined communities" (Gruzd, Wellman, and Takhteyev 2011). Yet, relationships are still constrained by geographic propinquity (Takhteyev, Gruzd, and Wellman 2012). Thus we now consider geographic aspects. Since highly reciprocal egonetworks indicate more intimate clusters, whereas networks with low reciprocity consist primarily of strangers, we hypothesize that:

Hypothesis 2 - Closely-knit networks are less geographically dispersed.

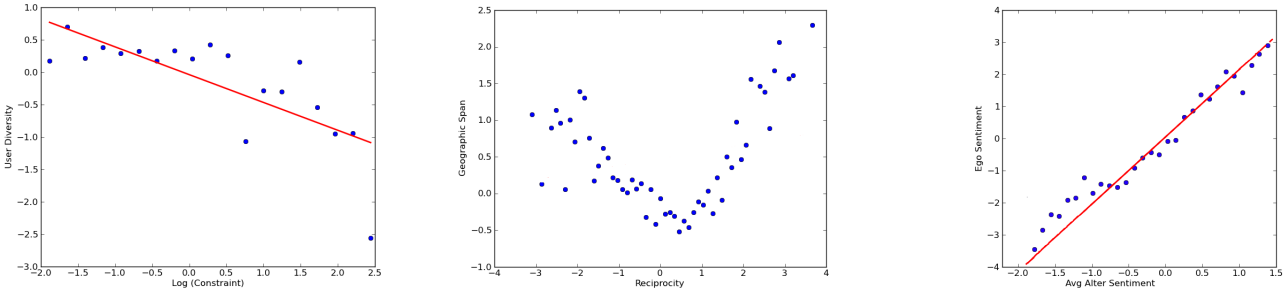
Geographic Span. To verify this hypothesis, the first step is to measure how geographically dispersed an egonetwork is. The dispersion can only be computed for those egonetworks where the locations of the ego and (a considerable part of) the alters are known. 157K users specified geographic locations (mostly city names) and converted these home locations into longitude-latitude pairs using the Yahoo! PlaceMaker API[1]. With this data at hand, we are now able to compute the geographic span of an ego's alters (Onnela et al. 2011) - higher span reflects more geographically dispersed networks, whilst low span indicates a geographically local network. In a way similar to (Onnela et al. 2011), we compute the geographic span of an egonetwork G as the average Euclidean distance of each alter from the ego $D_G = (1/n_G) \sum_{i \in V} \sqrt{(X_{\hat{v}} - x_i)^2 + (Y_{\hat{v}} - y_i)^2}$, where D_G is the geographic span of egonetwork G; n_G is the number of alters in the egonetwork for whom a location can be inferred; $(X_{\hat{v}}, Y_{\hat{v}})$ is the location of the ego; (x_i, y_i) is the location of the i^{th} alter; V is the set of all alters whose location is available; and the sign '-' is not the arithmetic subtraction but is the difference operator for angular measurements.

We computed the distribution of the geographic span for the 784 egonetworks for which location information was available. The distribution is log-normal: few users have a very nucleated span, most of them are fairly distributed to some degree, with a few being exceptionally large. By then looking at the four different versions of egonetwork we constructed, we find that the stronger the ties in a network, the lower the network's geographic span - the geometric average of geographic span is 23.5 for *reciprocal* egonetworks, 14.4 for *1-way interaction* egonetworks, and 7.84 for *2-way interaction* egonetworks.

Hypothesis testing. We next compute the correlation coefficients between network measures and geographic span, and find that users with geographically local networks tend to:

- Have more socially constrained networks. Higher network constraints are associated with more intimate networks, which in turn, as one expects, tend to be geographically local (the corresponding correlations are negative). Also, the stronger the ties, the more important geography: the correlation between network constraint and geography goes

[1] http://developer.yahoo.com/geo/placemaker/



(a) Diversity vs. Constraint (b) Geo Span vs. Reciprocity (c) Sentiment: Ego vs. Alters

Figure 3: Plot of binned values for: **(a)** normalized topical diversity against (log-transformed) network constraint; **(b)** normalized geographic span against normalized network reciprocity; and **(c)** average sentiment of an ego and that of the ego's alters.

Geo Span	**Network Constraint** (no access to structural holes)			**Reciprocity**	**Simmelian**	**Following**	**Followers**	**Status**
	Reciprocal	1-way Msg	2-way Msg					
Mutual	**-0.11**			0.02, any r **0.40**, $r > 0$ **-0.29**, $r < 0$	0.10	-0.01	-0.06	-0.08
Unicast		**-0.19**		0.05, any r **0.32**, $r > 0$ **-0.24**, $r < 0$	0.02	0.11	0.08	-0.01
Bicast			**-0.21**	0.06, any r **0.40**, $r > 0$ **-0.18**, $r < 0$	0.02	0.00	0.02	0.03

Table 2: Correlation coefficients r between network properties and geographic span of egonetworks. Highlighted are those results that are statistically significant (p-values < 0.01).

from -0.11 for *original* egonetworks to -0.21 for the more "social" *2-way interaction* egonetworks (Table 2).

- Not to be characterized by reciprocal networks. We found no correlation between geographic span and network reciprocity. However, by plotting geographic span against network reciprocity, a parabolic shape is visible (Figure 3b), showing that there are two linear relationships of opposite signs, which meet at the average value of reciprocity (i.e., at the normalized reciprocity $\tilde{r} = 0$). This means that networks that deviate from the typical (average) value of reciprocity (the minority) are geographically dispersed, while those with typical reciprocity (the majority) are local (minimum geographic span). This suggests that a minority of profiles shows anomalous reciprocity levels. We will see how this can be exploited for detecting accounts that are unlikely to belong to real people (Section 7).

6 Networks and Emotions

We have seen that there are strong relationships between network measures and the two properties of what people say (topical diversity) and where people are (geographic span). Another property that has been recently studied is how people share emotions online (Kivran-Swaine and Naaman 2011). Back in 1986, McMillan and Chavis (1986) posited that a "shared emotional connection" is integral for an individual to feel a sense of belonging to a community (McMillan and Chavis 1986). The prevailing belief amongst psychologists is that social sharing of emotion is a cathartic and bonding act, and that the extent to which emotions are shared within a social network is associated with the strength of the ties within that network (McMillan and Chavis 1986; Granovetter 1973; Kivran-Swaine and Naaman 2011). In other words, individuals are more likely to share emotions with their close friends than with strangers. Therefore, we would expect that users who are more inclined to share their emotions have more intimate networks – featuring a high proportion of mutual or Simmelian ties. So our next working hypothesis is:

Hypothesis 3 - Tweeters are more likely to share their emotions in more closely-knit ego-networks.

Another aspect associated with emotions is whether and, if so, how they are clustered in a network. Fowler and Christakis undertook a study of emotional clustering amongst participants in the Framingham Heart Study in Framingham (Fowler and Christakis 2008). Individuals were assessed according to the Center for Epidemiological Studies depression scale (CES-D) in which participants were asked how often they experienced certain positive feelings during the previous week, such as "I was happy" or "I enjoyed life". The researchers found that "People who are surrounded by many happy people and those who are central in the network are more likely

	Network Constraint (no access to structural holes)				Reciprocity	Simmelian	Following	Followers	Status
	Original	Reciprocal	1-way Msg	2-way Msg					
Emotion									
LIWC	0.05	0.07	**0.16**	**0.18**	0.02	-0.07	-0.07	0.02	0.06
MaxEnt	0.06	0.07	**0.24**	**0.27**	0.01	-0.13	-0.07	0.02	0.02
Sentiment									
LIWC	**-0.23**	**-0.25**	-0.06	-0.05	**0.12**	-0.04	0.23	0.15	0.02
MaxEnt	**-0.18**	**-0.24**	**-0.19**	**-0.12**	**0.13**	0.05	0.19	0.15	0.06

Table 3: Correlation coefficients r between network properties and expression of emotion and sentiment in tweets. Highlighted are those results that are statistically significant (p-values < 0.01).

to become happy in the future". Their findings are twofold: (a) that happy individuals tend to cluster together; and (b) that individuals *spread* happiness. To test whether the first finding (i.e., clustering) also applies to Twitter, we test whether an ego's sentiment and the average sentiment of the ego's alters are correlated. So a further hypothesis about emotions is:

Hypothesis 4 - Users who express positive (negative) emotions have alters who do likewise.

To put the two hypotheses related to test, we need to classify the sentiment of tweets first.

Sentiment Classification. We measure the sentiment of a profile using two classifiers: Word Count and Maximum Entropy. *Word Count* relies on a dictionary called "Linguistic Inquiry Word Count". *LIWC* is a standard dictionary of 2,300 English words that capture 80% of the words used in everyday conversations and reflect people's emotional and cognitive perceptions. After removing stop-words from tweets, we count, for each profile, the number of words that are positive and those that are negative (words matching the two categories of 'positive emotions' and 'negative emotions' as defined in *LIWC*) and aggregate both counts to produce the *LIWC* score, which is similar to the score proposed by Kramer (Kramer 2010):

$$Sentiment_i^{WC} = \frac{p_i - \mu_p}{\sigma_p} - \frac{n_i - \mu_n}{\sigma_n} \quad (2)$$

where p_i (n_i) is the fraction of positive (negative) words for user i; μ_p (μ_n) is the average fraction of positive (negative) words over all users; and σ_p (σ_n) is the corresponding standard deviation. The normalization using means and standard deviations accounts for the unbalanced distribution of positive and negative words of the English language (Kramer 2010). *Maximum Entropy*, instead, is a machine learning technique that has been proven to be effective in a number of natural language processing applications, including sentiment classification of tweets (Barbosa and Feng 2010). We use *MaxEnt* to classify tweets and then compute a profile's sentiment using, again, formula (2).

Effectiveness of classifiers. Having the two classifiers at hand, we previously measured how well they performed (Quercia et al. 2011; 2012). Upon 10-fold cross validation, we found that the two classifiers showed very similar tweet-level *accuracy* upon the tweets they were able to classify (precision is around 66%) but exhibited different *recall*, in that, *LIWC* left more tweets unclassified than what *MaxEnt* did (recall was 38% for *LIWC* and 68% for *MaxEnt*). However, these results are for single tweets. At profile level, the two classifiers performed very similarly instead, and their classifications were strongly correlated (Pearson correlation coefficient of $r = .73$): each profile, on average, was considered to be positive/negative to a very similar extent by both classifiers.

Emotion Words. In addition to the sentiment of a user's tweets, we are interested in the extent to which a user's tweets are emotionally charged (regardless of positivity or negativity). The emotion score for a user i is defined as:

$$e_i = \frac{|P_i| + |N_i|}{|T_i|}$$

that is, the proportion of words in the user's tweets that are positive ($|P_i|/|T_i|$) or negative ($|N_i|/|T_i|$) over the total number of tweets $|T_i|$. The distribution of emotion words is normal.

Hypotheses testing. From Table 3, one sees that *LIWC* and *MaxEnt* classifications produce the same correlations. The presence of emotion words does not correlate with any network metric other than network constraint, and it does so only for networks with strong ties, which are formed by looking at who exchanges messages with whom (*1-way interaction* and *2-way interaction* networks). In those cases, emotion words tend to be expressed in more constrained and intimate networks. As for sentiment, a more detailed picture emerges. We find that negative emotions are expressed in less constrained networks. This result complements two recent findings in Twitter. The first is that Tweeters have a greater tendency to share emotion in sparser networks (Kivran-Swaine and Naaman 2011), and the second is that Tweeters who are influential tend to freely express negative emotions (Quercia et al. 2011). Here we see that also those who have access to brokerage opportunities tend to express negative emotions.

To test our second hypothesis (i.e., to determine whether the clustering of mood within egonetworks exists), we plot

one's sentiment versus the average sentiment of one's alters (Figure 3c) and find fairly conclusively that there is indeed a clustering of emotions, with correlation coefficients as high as $r = 0.45$.

7 Conclusion

We have presented a number of insights that make it possible to compare social dynamics in Twitter to those in physical communities. Much like the real-world, those who have brokerage opportunities in Twitter tend to cover diverse topics; the majority of users have geographically-constrained networks; and "happy" ("sad") users do cluster together. These findings are not surprising as social media sites have been built around people (not around content), and the social behavior we have evolved over thousands of years is what drives the actions in those sites. For the first time, we can accurately capture social interaction, and many of our theories can now be quantitatively tested.

Understanding social dynamics in Twitter benefits not only researchers who are interested in social dynamics but also has practical implications:

(a) Quantified Self. One could imagine user interfaces that show how one's profile approaches (or deviates from) the expected values of topical diversity, geographic span, and expression of emotions, making users aware of their actions. This is important because our conscious brains are not designed to process huge amounts of information and, as such, most of our behaviour is driven by our non-conscious brain (Kahneman 2011).

(b) Spam Detection. Based on network reciprocity levels, one could identify which users are unlikely to be real users. To test this assertion, we take the unfiltered 258,895 Twitter profiles we have crawled and gathered each profile's TrstQuotient score provided by Infochimps.com (low value are indicative of abusive or spam accounts, high values are associated with real people). We formulate a task of predicting spam accounts as a binary classification problem, where the response variable is whether a profile is in the bottom or top quartile of TrstQuotient scores. After excluding the two middle quartiles, we are left with a balanced sample (the response variable is split 50-50), and the accuracy of a random prediction model would be 50%. Using a logistic regression to perform the binary classification on input of geographic span, we are able to correctly classify 87% of accounts.

(c) Marketing. It is tempting to think that a user connects to a very diverse set of people and that Twitter allows us to connect to thousands of individuals. A different picture has however emerged: similar users connect with each other (e.g., users who are connected tend to express emotions in the same way). Identifying a user's restricted social circle will move marketing campaigns away from simply segmenting by demographics and psycho-graphics and interrupting people to grab their attention (the dominant form of marketing for the last 50 years). It will move them towards segmenting by social network structure (e.g., social brokerage) and supporting conversations of small social circles about businesses. When it comes to spreading ideas, one needs to target users' closest ties who hold a disproportionate amount of influence.

The main limitation of this study (which calls for further work) is that our results do not speak of causality and are based on the last 200 tweets of each profile. To fix that, one could crawl Twitter over multiple time intervals, use a cross-lag analysis to examine potential causal relationships, and study how topical diversity, geographic properties, expression of emotions, and network properties evolve over time.

Acknowledgment. We thank Jonathan Ellis for collecting part of the data and thank EPSRC for its financial support through the Horizon Digital Economy Research grant (EP/G065802/1).

References

Anderson, B. R. 1983. *Imagined communities: reflections on the origin and spread of nationalism*. Verso, London.

Aral, S., and Alstyne, M. W. V. 2011. The Diversity-Bandwidth tradeoff. *American Journal of Sociology* 117(1):90–171.

Barbosa, L., and Feng, J. 2010. Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the* 23^{rd} *COLING*.

Burt, R. S. 1992. *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.

Burt, R. S. 2005. *Brokerage and Closure: An Introduction to Social Capital*. Oxford University Press.

Butcher, M. 2009. London is the capital of Twitter, says founder. TechCrunch Europe.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of International AAAI Conference on Weblogs and Social (ICWSM)*.

Fowler, J. H., and Christakis, N. A. 2008. Dynamic spread of happiness in a large social network. *BMJ* 337.

Granovetter, M. S. 1973. The Strength of Weak Ties. *The American Journal of Sociology* 78(6):1360–1380.

Gruzd, A.; Wellman, B.; and Takhteyev, Y. 2011. Imagining Twitter as an Imagined Community. *American Behavioral Scientist, Special issue on Imagined Communities*.

Hansen, M. 1999. The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly*.

Huberman, B. A.; Romero, D. M.; and Wu, F. 2008. Social Networks that Matter: Twitter Under the Microscope. *First Monday*.

Kahneman, D. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kivran-Swaine, F., and Naaman, M. 2011. Network properties and social sharing of emotions in social awareness streams. In *Proceedings of the ACM CSCW*, 379–382.

Kramer, A. 2010. An unobtrusive behavioral model of "Gross National Happiness". In *Proceedings of the* 28^{th} *ACM CHI*.

McMillan, D. W., and Chavis, D. M. 1986. Sense of community: A definition and theory. *Journal of Community Psychology* 14(1):6–23.

Onnela, J.-P.; Arbesman, S.; González, M. C.; Barabási, A.-L.; and Christakis, N. A. 2011. Geographic Constraints on Social Network Groups. *PLoS ONE* 6(4):e16939+.

Quercia, D.; Askham, H.; and Crowcroft, J. 2012. TweetLabel: A supervised topic model for assigning topics to Twitter profiles. In *Proceedings of the* 4^{th} *ACM Web Science*.

Quercia, D.; Ellis, J.; Capra, L.; and Crowcroft, J. 2011. In the Mood for Being Influential on Twitter. In *Proceedings of the* 3^{rd} *IEEE SocialCom*.

Quercia, D.; Ellis, J.; Capra, L.; and Crowcroft, J. 2012. Tracking "gross community happiness" from tweets. In *Proceedings of the* 15^{th} *ACM CSCW*.

Takhteyev, Y.; Gruzd, A.; and Wellman, B. 2012. Geography of twitter networks. *Social Networks* 34:73–81.

Tortoriello, M., and Krackhardt, D. 2010. Activating cross-boundary knowledge: the role of simmelian ties in the generation of innovations. *Academy of Management*.